DOCUMENT RESUME

ED 243 917 TM 840 215

AUTHOR Yen, Wendy M.
TITLE The Choice of Scale for Educational Measurement: An IRT Perspective.
PUB DATE Apr 84
NOTE 37p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Viewpoints (120)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Tests; Educational Research; *Latent Trait Theory; Mathematical Models; *Scaling; Scoring; Testing; *Test Interpretation
IDENTIFIERS *Equal Interval Scales; *Thurstone Scales

ABSTRACT

Two of the most popular methods for obtaining equal-interval scales for educational measurement are discussed: Thurstone's method and Item Response Theory (IRT). Between-grade growth on these scales is compared; while unstandardized differences show different trends for the two scales, standardized differences that take standard deviations into account show very similar growth patterns for the two scales. IRT is used with simulated data to demonstrate that comparisons of growth at different percentiles is not appropriate when observed scores, which contain error of measurement, are the basis of the percentiles. Another IRT simulation demonstrates that normalized observed scores will not necessarily be linearly related to underlying normally distributed traits when error variance affects the observed score distribution. When achievement is multidimensional and changes with age, it is not possible to obtain an equal-interval scale that extends over test levels. The outcome of the use of unidimensional IRT with multidimensional tests is discussed, and several alternative IRT scales are outlined. Finally, a brief discussion of statistical analyses of achievement scales is presented. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED243917

The Choice of Scale for Educational Measurement: An IRT Perspective

Wendy M. Yen

CTB/McGraw-Hill

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. M. Yen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April 1984

TM 840 215

Abstract

Two of the most popular methods for obtaining equal-interval scales for educational measurement are discussed: Thurstone's method and Item Response Theory (IRT). Between-grade growth on these scales is compared; while unstandardized differences show different trends for the two scales, standardized differences that take standard deviations into account show very similar growth patterns for the two scales. IRT is used with simulated data to demonstrate that comparisons of growth at different percentiles is not appropriate when observed scores, which contain error of measurement, are the basis of the percentiles. Another IRT simulation demonstrates that normalized observed scores will not necessarily be linearly related to underlying normally distributed traits when error variance affects the observed score distribution. When achievement is multidimensional and changes with age, it is not possible to obtain an equal-interval scale that extends over test levels. The outcome of the use of unidimensional IRT with multidimensional tests is discussed, and several alternative IRT scales are outlined. Finally, a brief discussion of statistical analyses of achievement scales is presented.

The Choice of Scale for Educational Measurement: An IRT Perspective

Many types of scales are available to measure educational achievement: for example, number-correct scores, percentiles, grade equivalents, and scale scores. Most of the scales are monotonic transformations of each other; they rank examinees in the same way, but they differ in terms of scale units. This paper focuses on "equal-interval" scales that are available for measuring change in individual examinees' achievement within and across test levels. Insights gained through the use of Item Response Theory (IRT) in examining the properties of these scales will be examined.

In the terminology of measurement theory, the legitimate method of producing an equal-interval scale is through "fundamental" measurement (Lingoes, 1979; Torgerson, 1958), sometimes called "representational" measurement (Dawes, 1972). Fundamental measurement requires the use of a scaling model with which predictions about the internal relationships among a set of measurements are made and tested. To satisfy the criterion of fundamental measurement, the scale must be subject to verification or falsification via data. The type of admissible transformation of the scale in the model that will not alter the model's predictions defines the level of measurement produced. For example, if only a linear transformation of the scale will preserve the model's predictions, an interval level of measurement is reached, and an equal-interval scale is produced.

In educational measurement there have been two major methods used to produce equal-interval scales through fundamental measurement: Thurstonian scaling and IRT. Thurstonian scaling was first introduced to a widely-used, major achievement battery with the Comprehensive Tests of Basic Skills, Form Q, in 1968 (CTB/McGraw-Hill, 1968). Since then this method has become the most widely used method of obtaining equal-interval scales for educational achievement tests. The Thurstonian method involves normalizing number-correct scores within two or more groups of examinees, each of which is hypothesized to have a normal distribution of the trait in question. The test of the internal consistency of the model comes from a comparison of the results of the normalizations; if there is a linear relationship between the normalized scores at specified percentiles in each group with the normalized scores in the other groups, the model predictions are verified. In practice the Thurstonian method is typically used in conjunction with equipercentile equating: the scores for each of two test forms or levels are normalized for equivalent groups of examinees and a linear equipercentile equating is applied to these normalized scores. Thus, in practice the most common method of vertically scaling many test levels combines normalization of scores by grade within test level and equipercentile equating across test levels. This procedure that combines test scaling and equating will be called the Thurstonian method.

IRT provides an alternative method of establishing an equal-interval scale through fundamental measurement. Instead of making predictions about distributions of total scores, IRT makes predictions about the probability that examinees at different trait (scale) levels will correctly answer each item. In comparison with the Thurstonian method IRT produces much more precise and varied predictions about the relationships among tests, items, and examinees. When these predictions are accurate, the IRT model produces an equal-interval scale according to the requirements of fundamental measurement.

For many years Thurstonian scaling was the only method used for scaling educational achievement tests. The properties of the Thurstonian scale came to have psychological or educational significance. For example, because the Thurstonian scale tended to produce scores that had increasing variance as grade increased, people began to believe that "achievement" necessarily increased in variance as grade increased. This belief occurred despite the fact that a precise definition of "achievement" did not exist apart from the specification of the items in a test. There was no careful hypothesis building or model testing about the psychological nature of the "achievement" trait and how it developed or was influenced by various circumstances.

This acceptance of the Thurstonian scale as being linearly related to "achievement" was not effectively challenged for many years, largely because there was no strong alternative

scaling method. Grade equivalents had been promoted by some as an appropriate equal-interval scale for educational measurement. However, grade equivalents did not qualify as an equal-interval scale according to measurement theory because no model predictions or verifications were involved in their establishment; they were defined by fiat, and as such there was no way to establish them as fundamental measurements. Also, interpretive problems with grade equivalents were great enough that they never became widely accepted as defining an equal-interval scale.

With the introduction and widespread use of IRT an important alternative to Thurstonian scaling was available. It soon became apparent that the IRT models tended to produce scalings that differed from the Thurstonian scalings. Given that the two methods have no assumptions or procedures in common, it should not have been a surprise that their results might differ. However it has been a surprise, and people have begun taking sides, stating that one method is necessarily right and the other is wrong.

Those who claim to know that a particular method is right must believe that achievement is unidimensional, that it can be linearly ordered, and that students can be located in this linear ordering independently of performance on a particular achievement test. To my knowledge, no one has succeeded in identifying or defining a linearly ordered psychological trait in educational achievement, and no one has demonstrated that a

particular measurement scale is linearly related to such a trait.

For many years educators have asked what have appeared to be reasonable and important questions about their students' achievement: How much does this program improve achievement relative to that program? How much are my low-achieving students learning relative to my high-achieving students? Are my third graders learning at the same rate as my tenth graders? Is this student learning as much in sixth grade as he or she learned in fifth grade? Educators have asked these questions, and test developers have offered them a way to get answers.

All these questions about change in learning or achievement have at least a tacit assumption about the existence of linearly ordered achievement traits such as reading vocabulary level or mathematics computation skill, and answering them requires the use of an equal-interval scale. Test developers have offered what they believed to be appropriate scales to answer these questions, and many test users have believed that they were getting appropriate answers to their questions. The recent opportunity to choose between different putative equal-interval scales and the possibility that they would provide different answers to the questions, leads us not only to ask, Which scale is right?, and, Is either scale right?, but also, Do we have a way to determine which scale is right? and even, Do the educators' questions have meaning?

The main obstacle to the establishment of truly (verifiably) equal-interval achievement scales is the fact that achievement is multidimensional. At any given time during learning there are many factors in the environment and in the student that influence the student's behavior. Further, the nature of what is being learned is constantly changing. The student is learning the names of letters of the alphabet one day, associating sounds to these letters another day, and attaching meaning to groups of letters a third day. Does it show more growth to move from adding whole numbers without carrying to adding whole numbers with carrying or to move from adding fractions to multiplying fractions?

No one has specified all the events that occur during learning, and no one has specified how all the parts of learning combine to move a student one unit further along some hypothetical achievement scale. The task of fully specifying learning or achievement is impossible, and the educators' questions cannot be answered.

Despite the fact that the educator's questions cannot really be answered, I do not expect educators to stop asking their questions or for test developers to stop offering ways to answer them. It is important for educators and test developers to acknowledge that truly equal-interval scales for educational measurement are not possible. It is necessary for those offering scales to obtain and provide as much information as possible about the properties of their scales, and it is

important for those using test scores as data in answering questions to analyze this data in such a manner that the lack of equal-intervals is least likely to distort results.

The following section of this paper attempts to shed light on the properties of the Thurstonian and IRT scales. A variety of questions are asked of the scales, and most of the analyses are pursued from an IRT perspective. The IRT perspective is favored because, in contrast to the Thurstonian model, it is rich in predictions and means of verifying them. Finally, a short section is presented that deals with some of the issues related to statistical analysis of test scores.

Some Relevant IRT Studies

Between-Grade Growth on Thurstonian and IRT Scales

Although studies of the Thurstonian and IRT scales have not been made across several levels of the same test, it is likely that such studies would show a nonlinear relationship between the two scales. The California Achievement Tests, Form C (CAT/C; CTB/McGraw-Hill, 1977) were vertically scaled with the Thurstonian method and normed in 1976-1977. The Comprehensive Tests of Basic Skills, Form U (CTBS/U; CTB/McGraw-Hill, 1981) were vertically scaled with IRT and normed in 1980-1981. Table 1 contains means and standard

Insert Table 1 about here

deviations of scale scores by grade for the Spring 1977 norm group for CAT/C and in the Spring 1981 norm group for CTBS/U for the Reading Vocabulary and Mathematics Computation tests. The standard deviations tend to increase with grade for CAT/C and decrease with grade for CTBS/U. Different standardization groups are being compared here. While it is possible that the differences in standard deviations reflect norm group or test differences, it is much more likely that they arise from differences in the Thurstonian and IRT scales. Standardized and unstandardized mean differences between grade means are in Table 2. The standardized differences are the unstandardized

Insert Table 2 about here

differences divided by the square root of the average within-grade variances for the two grades whose means are compared. For all four tests the unstandardized between grade growth tends to decrease as grade increases. This effect is more apparent for the CTBS/U tests than the CAT/C tests and is particularly noticeable for the Mathematics Computation test. However, the standardized differences are very similar for the CTBS/U and CAT/C tests, despite the fact they are based on different normative groups. In other words the standardized

differences lead to essentially the same conclusions, regardless of which scale is used. The differences between the two scales have little impact when standaridized differences are examined.

Comparing Growth at Different Percentiles

It is common for growth comparisons to be made between examinees at different percentiles. For example, it might be stated that for an examinee to stay at the 25-th percentile when going from grade 4.1 to grade 5.1 the examinee must grow X scale score units, while for an examinee to stay at the 75-th percentile he or she must grow Y scale score units. This section shows that it is not reasonable to calculate or compare change at selected percentiles for any test whose standard error of measurement is not constant for all scores.

Yen (1983) presents simulation data demonstrating that two tests that differ in difficulty level will produce different distributions of observed scores (traits) due to the tests' having different amounts of error variance in different parts of the scale. For example, a more difficult test will tend to have more error variance and therefore more observed score variance among low-scoring examinees than will an easier test. On the other hand, the easier test will have greater error and observed score variance among high-scoring examinees than will the more difficult test. Thus, even though scores on the two

tests are tau-equivalent (i.e., an examinee expects to obtain the same score on the average from the two tests), the examinee will expect to be at different percentiles with respect to the observed score distributions on the two tests.

A similar principle applies to examinees at different performance levels taking the same test. For example, consider students who are tested twice with the same test with one year intervening between testings. If all the students grow exactly the same amount, they are not expected to maintain the same percentiles with respect to the distributions of observed scores. This counter-intuitive result occurs because of the redistribution of the majority of the error variance between one testing and the other. Say, for example, that most of the error variance occurs among low-scoring examinees on the first testing because the test is somewhat difficult. On the second testing the low-scoring examinees move up to where the test is measuring well, and the high-scoring examinees move up beyond where the test is measuring well. In this example there would be a shift of the predominant amount of error variance from among low-scoring examinees on the first testing to among high-scoring examinees on the second testing. The observed score distribution would change its shape because of this shift in the location of the error variance, and even if every examinee's true score grew the same amount, the examinees' percentiles would change systematically from the first to the second testings.

A simulated example of this effect can be presented. This example is based on a 20-item simulated test described by Yen (1983). The three-parameter logistic model was used with the item difficulties (b_i values) set at the following values: -2.19, -1.88, -1.56, -1.24, -.99, -.88, -.78, -.67, -.57, -.46, -.36, -.25, -.15, -.04, .06, .16, .48, .79, 1.11, and 1.42. The item discriminations (a_i values) were all set at 1.0 and the item lower asymptotes (c_i values) were all set at .2. True traits (θ_k values) for 1000 simulees were generated from a normal distribution with a mean of -.5 and a standard deviation of .9 using the IMSL normal random deviate generator GGNML (IMSL, 1979). It was assumed that for the first testing the observed trait value ($\hat{\theta}_k$) for each simulee had a normal distribution with mean θ_k and variance $1/I(\theta_k,\hat{\theta})$, where $I(\theta,\hat{\theta})$ was the information function predicted by the three-parameter model for θ_k and the prespecified item parameters. For the second testing it was assumed that the trait values for each simulee had a normal distribution with mean $\theta_k + 1.0$ and variance $1/I(\theta_k+1, \hat{\theta})$. Using these assumptions two observed trait values were generated (using GGNML) for each simulee to simulate observed scores on the same test on different testing occasions, when all the simulees' true trait values had grown the same amount between testing occasions.

The mean and standard deviation of the observed trait values were -.04 and 1.05 for the first testing and .97 and 1.20 for the second testing. On the basis of the percentile

ranks of their observed scores (trait estimates) on the first testing the simulees were grouped into 20 cells with 50 simulees in each cell. For example, the first cell contained simulees with percentile ranks from 0 to 5. The same type of grouping was performed for the trait estimates for the second testing. The mean difference between trait estimates on the two testings was obtained by cell. These differences are estimates of the amount of growth for simulees at fixed percentiles. The true amount of growth (1.0) was subtracted from each of the estimated differences. Figure 1 displays the differences

---

Insert Figure 1 about here

---

between estimated and true growth by percentile. In this example, holding percentile fixed while examining growth produces a substantial overestimate of growth in the lowest percentiles, a moderate overestimate of growth in the highest percentiles, and a substantial underestimate of growth in the moderate percentiles. This local bias in the measurement of growth occurs despite the fact that overall mean difference between estimated and true growth was .01.

Figure 1 presents just one example of how misleading results can be obtained by comparing examinees at fixed percentiles. The amount of distortion produced by the procedure of examining growth for fixed percentiles will depend

upon the particular test used and its relationship to the distribution of trait values on the first and second testings.

Normalizing Score Distributions

The basic assumption of the Thurstonian scaling method is that the ability or trait being scaled is normally distributed in a particular population. For psychological traits it is virtually impossible to establish the validity of this assumption because of the difficulty described in the introductory section of defining the trait scale. For some traits a skewed or bimodal distribution might be reasonable (for example, see Bock & Kolakowski, 1973). In spite of this difficulty, let us assume that a particular trait is normally distributed in a particular population. If a distribution of observed scores measuring this trait is normalized, will the normalized scores necessarily be linearly related to the underlying trait? The considerations described in the preceding section with respect to distributions of measurment error arouse the suspicion that normalization of observed scores is not necessarily appropriate.

Another simulation using IRT can shed light on this issue. The generation of the observed scores for this simulation is described by Yen (in press). Briefly, a distribution of 1000 true trait values was generated from a normal distribution with a mean of 0 and a standard deviation of 1.0 using GGNML.

Number-correct scores on two 20-item tests and two 40-item tests were obtained for these simulees using the three-parameter IRT model. The two tests differ in difficulty and are called the Easy and Difficult tests.

If the Thurstonian model is reasonable, normalization of these number-correct scores should produce a scale linearly related to the true (underlying) trait scale. For each test the standardized normal score (z score) associated with each number-correct score was obtained by the procedure described in Allen and Yen (1979, pp. 163-164). On the basis of their true trait values the simulees were grouped into quintiles (five groups of 200 simulees each) and the mean z score and mean true trait value were obtained for each group. If the z scores and true trait values are linearly related, the differences between the mean z scores and mean true trait values will be a linear function of true trait value. Figure 2 displays the

---

Insert Figure 2 about here

---

differences for the 20-item tests, and Figure 3 shows the

Insert Figure 3 about here

---

differences for the 40-item tests. The normalized scores are not linearly related to the underlying trait, although the relationship is much more linear for the 40-item test than the

20-item test.

These results are explainable in terms of the distribution of error variance discussed before. For the Easy test there is a preponderance of error variance among high-scoring examinees. Such high-scoring examinees will be less extreme in terms of percentiles on the observed score distribution than on the true trait distribution, and there will be a particularly large negative difference between z scores and true trait values for these high-scoring examinees. Conversely, for the Difficult test, the preponderance of error variance among low-scoring examinees will produce particulary large positive differences between z scores and true trait values for low-scoring examinees. These deviations between z scores and true trait values are less extreme for the 40-item tests than the 20-item tests because there is less error variance on the 40-item tests.

Thus, it is not necessarily reasonable to expect normalized observed scores to be linearly related to underlying traits. Because of the effects of error variance, it would be expected that the relationship between normalized observed scores and true trait values would vary as a function of the standard error (or information) curve associated with a test.

Item Complexity and Scaling

Table 1 indicates that the differences between the IRT and Thurstonian scales apparently are greater for the Mathematics Computation tests than the Reading Vocabulary tests. Among the CTBS/U tests the Mathematics Computation and Mathematics Concepts and Applications tests show a greater decrease in standard deviation of scale scores as test level increases than any of the other tests. The mathematics tests also show a greater increase in item discrimination as test level increases than the other CTBS/U tests.

Yen (1984) has offered an explanation of what might be called the scale shrinkage effect. The Mathematics Computation items tend to increase in complexity as test level increases. While at lower test levels the students add and subtract whole numbers, at higher test levels multiplication and division of whole numbers and computations with decimals and fractions are introduced. The more difficult items are more complex in that they require knowledge of the simpler computation skills. For example, multiplication requires knowledge of addition. It appears that as test level increases, the Mathematics Computation items measure more complex skills and more traits.

In Yen (1984) the relationship is derived between the trait and item parameters of the unidimensional three-parameter logistic model used in scaling test data and the traits and item parameters of a two-dimensional logistic model that is

hypothesized to be generating the item responses. The unidimensional item parameters and trait appear to be weighted averages of the generating item parameters and traits, with the weights proportional to the underlying item discriminations on the two traits. In simplified terms, an item's discrimination will increase as the number of important traits influencing performance on the item increases. An examinee's trait estimate will be a weighted average of the important traits influencing performance on the predominant items. The variance of the estimated item difficulties and traits will decrease (i.e., the scale will "shrink") as the number of traits influencing the items increases if the correlation between the items' difficulties on the underlying dimensions is less than the correlation between the underlying traits. For example, multiplying items might be said to measure both adding and multiplying traits, and typically the items' difficulty with respect to addition has a weak relationship to the items' difficulty with respect to multiplication. However, the multiplying and addition traits tend to be fairly highly correlated over examinees. In such a situation, the unidimensional IRT model will tend to produce a scale that has higher discrminations and lower trait and item difficulty variances as item difficulty increases.

The unidimensional IRT model is selecting a scale that describes as consistently as possible performance on both

unidimensional and multidimensional items. Through weighted averages, the model is summarizing several dimensions with a single score. Weighted averages are a very common method of simplifying multidimensional data. Although there is always room to debate whether a particular set of weights is best, the IRT model appears to be handling multidimensional items in a reasonable manner, and the "scale shrinkage" appears to be an understandable result of applying a unidimensional procedure to multidimensional data.

In a situation where items change in complexity or dimensionality no equating method is strictly appropriate: tests measuring different traits cannot be equated. Because there is no clear criterion for the quality of "equating" of tests that measure different traits, it is not clear which equating method works best in such a situation. In some cases it may be possible to split multidimensional tests into unidimensional subsets that can be appropriately scaled; this situation would occur when one trait influences one set of items and another trait influences other items. However, when several traits influence single items, the test cannot be separated into unidimensional subsets.

If equating of multidimensional tests continues, it is important to examine more carefully the logical errors that such a practice can produce. It is also important to be more cautious in comparing results from tests that differ a great deal in difficulty or content. As a convenience, test

publishers produce scales that go from Kindergarten to 12-th grade. While the legitimate purpose of these scales may be to define correspondence between successive pairs of test levels, the existence of such a broad scale might lead test users to the illegitimate comparison of widely separated levels. Although the "equating" of similar tests may not lead to serious problems, comparing results from tests that differ greatly in difficulty or content is bound to produce meaningless results.

Choosing Among IRT Scales

There are several IRT models that can be used to scale tests. The most commonly-used ones are the one-parameter (or Rasch), the two-parameter, and the three-parameter models. These models differ in their assumptions. If a set of data meet the assumptions of all three models, the models will produce the same scalings; that is, under such circumstances the trait scales produced by the different models will be linearly related. However, if the data are appropriate for the three-parameter model and not the other models, the three methods will produce non-linearly related scales (Yen, 1981). In particular, if items have nonzero asymptotes and the two- or one-parameter model is applied to them, the θ scale will be affected in a predictable way. Thus, the choice of model can influence the scale, even among IRT models.

Lord (1975) presents another possible IRT scale: $\omega(\theta)$. The item characteristic function using this scale is as follows:

$$P_i^*(\omega(\theta)) = c_i + \frac{1-c_i}{1 + \exp[-1.7a_i\ (\omega^{-1}(\omega(\theta)) - b_i\ )]}. \quad (1)$$

$\omega(\theta)$ is a monotonic increasing transformation of the θ scale chosen so that item discriminations are uncorrelated with item difficulties. The ω^{-1} function undoes the $\omega(\theta)$ transformation so that $P_i^*(\omega(\theta)) = P_i(\theta)$, where $P_i(\theta)$ is the three-parameter logistic model.

Lord (1975) argues that there is no fundamental reason for choosing the θ scale over some monotonic increasing transformation of it, and one might as well use the $\omega(\theta)$ scale if this transformed scale has desirable properties. In a similar vein, Lord (1980) presents another transformation of the θ scale:

$$\theta^* = Ke^{k\theta}. \quad (2)$$

K and k are arbitrarily chosen constants. Lord (1980) also discusses the true score scale, which is another monotonic transformation of the θ scale:

$$\zeta = \frac{1}{n} \sum_{i=1}^{n} P_i(\theta). \quad (3)$$

Each of these scales has advantages and disadvantages. For example, one might see it as an advantage to have item difficultes and discriminations uncorrelated, which will occur with the $\omega(\theta)$ scale and which might occur with the θ^* scale. The true score scale produces much less statistical bias than the θ scale (Lord, 1983), and is more stable than the θ scale (Yen, in press). On the other hand, the $\omega(\theta)$ scale and the ζ scale are dependent on the particular items chosen to define the scale, and the constants for the θ^* scale are chosen arbitrarily. If one ζ scale is used for examinees over a broad span of abilities, ζ will be strongly positively skewed among low-ability examinees and strongly negatively skewed among high-ability examinees.

Thus, there are a variety of scales within IRT that one might use to measure achievement. Even if it is established that a particular model fits a set of data, there are an infinite number of monotonic transformations of the θ scale and item parameters of that model that can be made to produce the same predictions. At this time there is no clear superiority of one scale transformation over another; in fact there is no widely-accepted set of criteria for evaluating the scales. The derivation and discussion of such a set of criteria would be a valuable addition to the scaling field.

Statistical Analysis of Achievement Data

Given the complexity of educational achievement, scores on a test can only be a gross summary of multidimensional psychological events. The accuracy of the predictions of a scaling model can be examined, and in the terminology of traditional measurement theory, one can try to establish an equal-interval scale through fundamental measurement. However, as Lord has pointed out, a monotonic transformation can always be applied to the scale, the inverse of this transformation incorporated into the model, and the accuracy of the model predictions maintained. The result is that the best situation that will be produced, and even this best situation is unlikely, is that the test scale will be monotonically related to some important underlying achievement trait. In other words, if the scaler is lucky he or she will end up with an ordinal scale.

There has been a long controversy about the appropriateness of using statistics that involve the calculation of means and standard deviations with scales that have ordinal but not equal-interval properties (for example, see Burke, 1974). Adams, Fagot, and Robinson (1965) developed a theory of appropriate statistics in which they stated, "It is proposed that a statistic be regarded as appropriate relative to statements made about it in case the truths of these statements are invariant under permissible transformations of the

measurement scale" (p. 99). This statement is highly relevant to the analysis of educational achievment data. The researchers or educators who are analyzing test data are ultimately concerned with the truth of the conclusions that they reach. If they conclude that one educational program improves achievement more than another program, they want this conclusion to be true regardless of the particular scale chosen.

The Adams, Fagot, and Robinson theory is predicated on the determination of admissible transformations. In the best situation available to educational researchers, the admissible transformation is a monotonic one. Under these conditions, only a statistic that is invariant under monotonic transformations can produce meaningful results. Many statistical procedures are available that satisfy this requirement; for example, Spencer (1983) recently described a set of such procedures that may be useful for educational research.

Even if seemingly appropriate statistical procedures are used, the researcher is not guaranteed meaningful results; the ultimate criterion for the value of the analysis is the truth of the statements made as a result of the analysis. Test scores can only summarize complicated multidimensional events,and the inherent meaninglessness of comparing psychological growth for students at widely different achievement levels must always be kept in mind.

Conclusions

In educational measurement there are many schools of thought about the equal-interval scaling issue. Some measurement specialists define a particular scale (for example, the grade equivalent scale) as being equal-interval because it appears reasonable to them to do so. Some people follow the fundamental measurement principle as the means of verifying a scale's properties, and others believe that the idea of an equal-interval scale of educational achievement is necessarily meaningless. It is likely that there will be many schools of thought on this issue for many years to come.

Whatever one's position, it is necessary to examine as fully as possible the properties of the scale being used. One clear advantage of the model-based methods such as IRT is that they are rich in detailed predictions about scale properties; this paper presented several examples of how an IRT model could be used to examine the implications and appropriateness of various statistical procedures. It should be noted that although an IRT model was used to show the problems that can be associated with Thurstonian scaling or with measuring change at fixed percentiles, these problems do not arise from the IRT model. The problems arise from ignoring the effects of error variance in observed scores. The IRT model is a very useful tool for demonstrating such problems, and other models suitable for such demonstrations might be found.

It is important not to accept cosmetic scale properties as evidence of good measurement. Observed score distributions can be smoothed until they shine and made to take on the appearance of any of a number of pleasing characterisics that conform to preconceptions about good measurement; however, such an appearance does not in any way mean that good measurement has taken place or that the most appropriate scale for educational measurement has been produced.

References

Adams, E.W., Fagot, R.F., & Robinson, R.E. (1965). A theory of appropriate statistics. Psychometrika, 30, 99-127.

Allen, M.J., & Yen, W.M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole.

Bock, R.D., & Kolakowski, D. (1973). Further evidence of sex-linked major-gene influence on human spatial visualizing ability. American Journal of Human Genetics, 25, 1-14.

Burke, C.J. (1974). Measurement scales and statistical models. In G.M. Maranell (Ed.), Scaling: A Sourcebook for behavioral scientists (pp. 42-55). Chicago,IL: Aldine Publishing Co.

California Achievement Tests, Form C. (1977). Monterey, CA: CTB/McGraw-Hill.

Comprehensive Tests of Basic Skills, Form Q. (1968). Monterey, CA: CTB/McGraw-Hill.

Comprehensive Tests of Basic Skills, Form U. (1981). Monterey, CA: CTB/McGraw-Hill.

Dawes, R.M. (1972). Fundamentals of attitude measurement. New York: John Wiley & Sons.

IMSL Library (7th Ed.). (1979). Houston, TX: International Mathematical and Statistical Libraries.

Lingoes, J.C. (1979). Measurement. In J.C. Lingoes, E.E. Roskam, & I. Borg (Eds.), Geometric representations of relational data: Readings in multidimensional scaling (pp. 1-43). Ann Arbor, MI: Mathesis Press.

Lord, F.M. (1975). The "ability" scale in item characteristic curve theory. Psychometrika, 40, 205-217.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. New York: Lawrence Erlbaum Associates.

Lord, F.M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. Psychometrika, 48, 233-245.

Spencer, B.D. (1983). On interpreting test scores as social indicators: Statistical considerations. Journal of Educational Measurement, 20, 317-333.

Torgerson, W.S. (1958). Theory and methods of scaling. New York, John Wiley & Sons.

Yen, W.M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5, 245-262.

Yen, W.M. (1983). Tau equivalence and equipercentile equating. Psychometrika, 48, 353-369.

Yen, W.M. (1984). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. Manuscript submitted for publication.

Yen, W.M. (in press). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. Journal of Educational Measurement.

Author Note

I would like to thank George R. Burket for many insightful discussions about scaling issues.

Table 1

Means and Standard Deviations of Scale Scores for the Spring of Each Grade for CAT/C (1977) and CTBS/U (1981)

	Reading Vocabulary				Mathematics Computation			
	CAT/C		CTBS/U		CAT/C		CTBS/U	
Grade	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
1	331	40	488	85	276	32	390	158
2	377	48	579	78	327	36	576	77
3	412	52	622	65	384	43	643	44
4	449	58	652	60	425	52	676	35
5	475	62	678	59	462	57	699	24
6	501	68	697	59	495	66	713	20
7	523	72	711	57	526	78	721	23
8	552	80	724	54	566	86	728	23
9	571	84	741	52	585	87	736	17
10	599	92	758	52	600	90	739	16
11	623	95	768	53	614	95	741	18
12	637	100	773	55	619	99	741	20

Table 2

Mean Spring Between-Grade Scale Score Growth

for CAT/C (1977) and CTBS/U (1981)

	Reading Vocabulary				Mathematics Computation			
	Unstandard.		Standard.		Unstandard.		Standard.	
Grade	CAT	CTBS	CAT	CTBS	CAT	CTBS	CAT	CTBS
2-1	46	91	1.05	1.12	51	186	1.50	1.50
3-2	25	43	.70	.60	57	67	1.44	1.07
4-3	37	30	.67	.48	41	33	.86	.83
5-4	26	26	.43	.44	37	23	.68	.77
6-5	26	19	.40	.32	33	14	.54	.63
7-6	22	14	.31	.24	31	6	.43	.37
8-7	29	13	.38	.23	40	7	.49	.30
9-8	19	17	.23	.32	19	8	.22	.40
10-9	28	17	.32	.33	15	3	.17	.18
11-10	24	10	.26	.19	14	2	.15	.12
12-11	14	5	.14	.09	5	0	.05	.00

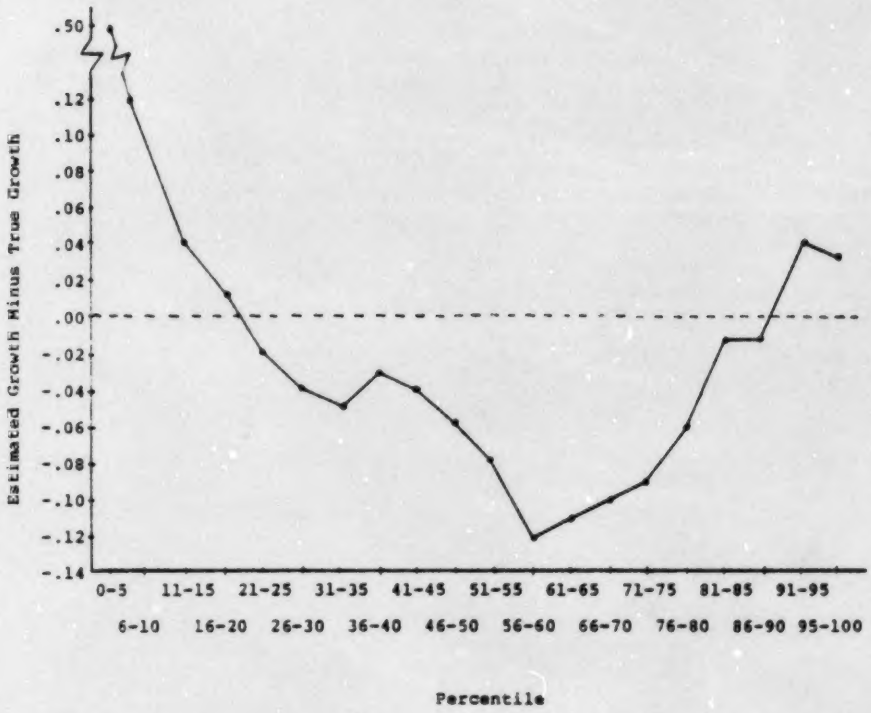

Figure 1. Estimated minus true growth as a function of percentile.

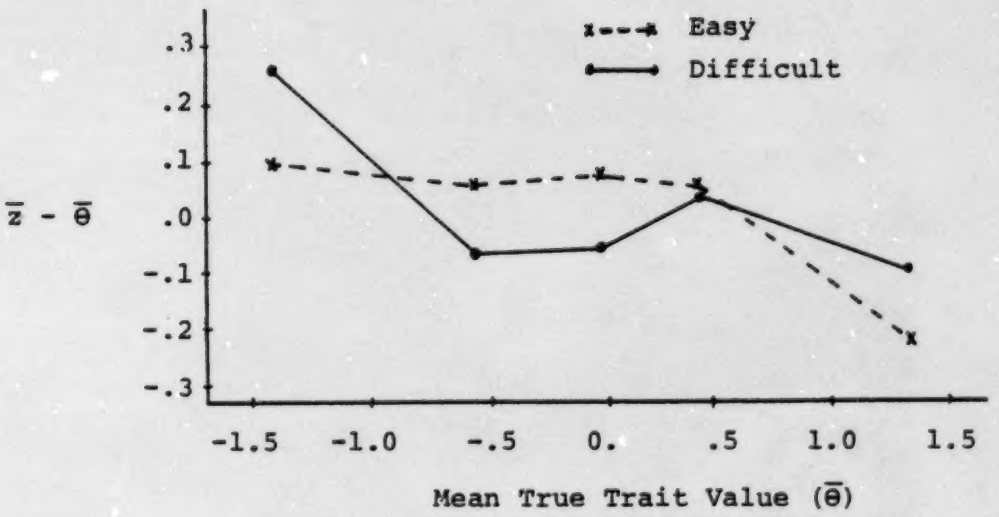

Figure 2. Mean z score minus true trait value as a function of true trait value for the 20-item tests.

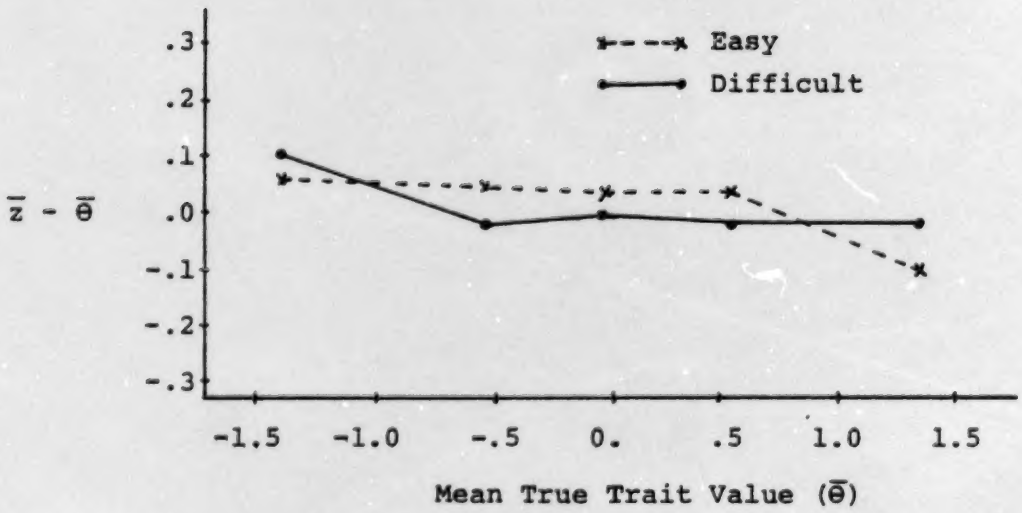

Figure 3. Mean z score minus true trait value as a function of true trait value for the 40-item tests.